Convex Optimization
CMU-10725

Independent Component Analysis
(and matrix differentials)

Barnabás Póczos & Ryan Tibshirani

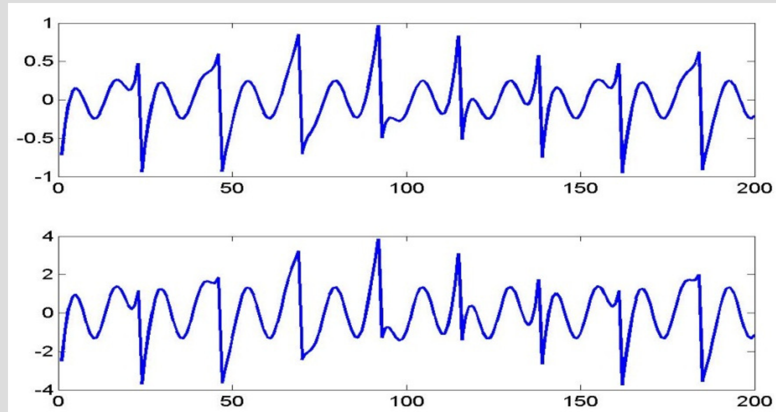
MACHINE LEARNING DEPARTMENT
Carnegie Mellon.
School of Computer Science

Independent Component Analysis

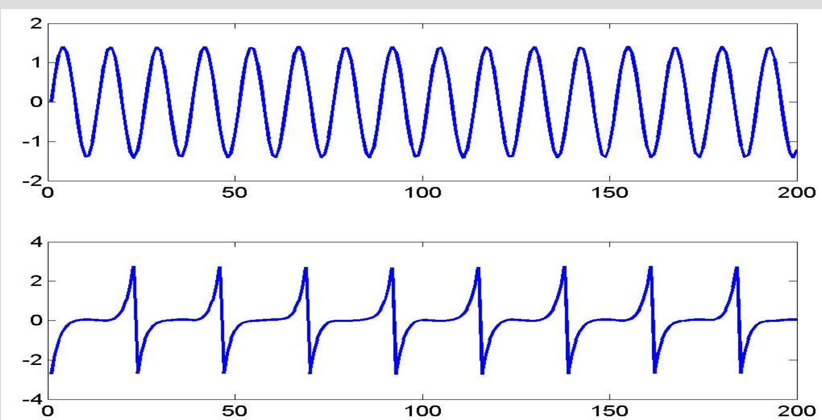
Independent Component Analysis

$$
\begin{aligned}
x_1(t) &= a_{11}s_1(t) + a_{12}s_2(t) \\
x_2(t) &= a_{21}s_1(t) + a_{22}s_2(t)
\end{aligned}
$$

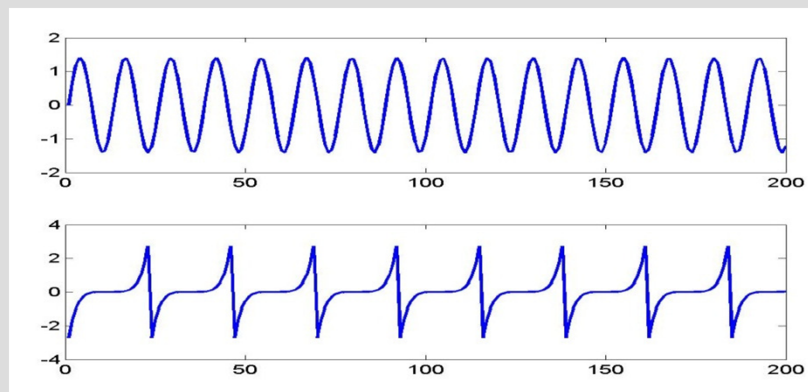
Model



Observations (Mixtures)



ICA estimated signals



original signals

Independent Component Analysys

Model

$$\begin{aligned} x_1(t) &= a_{11}s_1(t) + a_{12}s_2(t) \\ x_2(t) &= a_{21}s_1(t) + a_{22}s_2(t) \end{aligned}$$

We observe

$$\begin{pmatrix} x_1(1) \\ x_2(1) \end{pmatrix}, \begin{pmatrix} x_1(2) \\ x_2(2) \end{pmatrix}, \ldots, \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix}$$

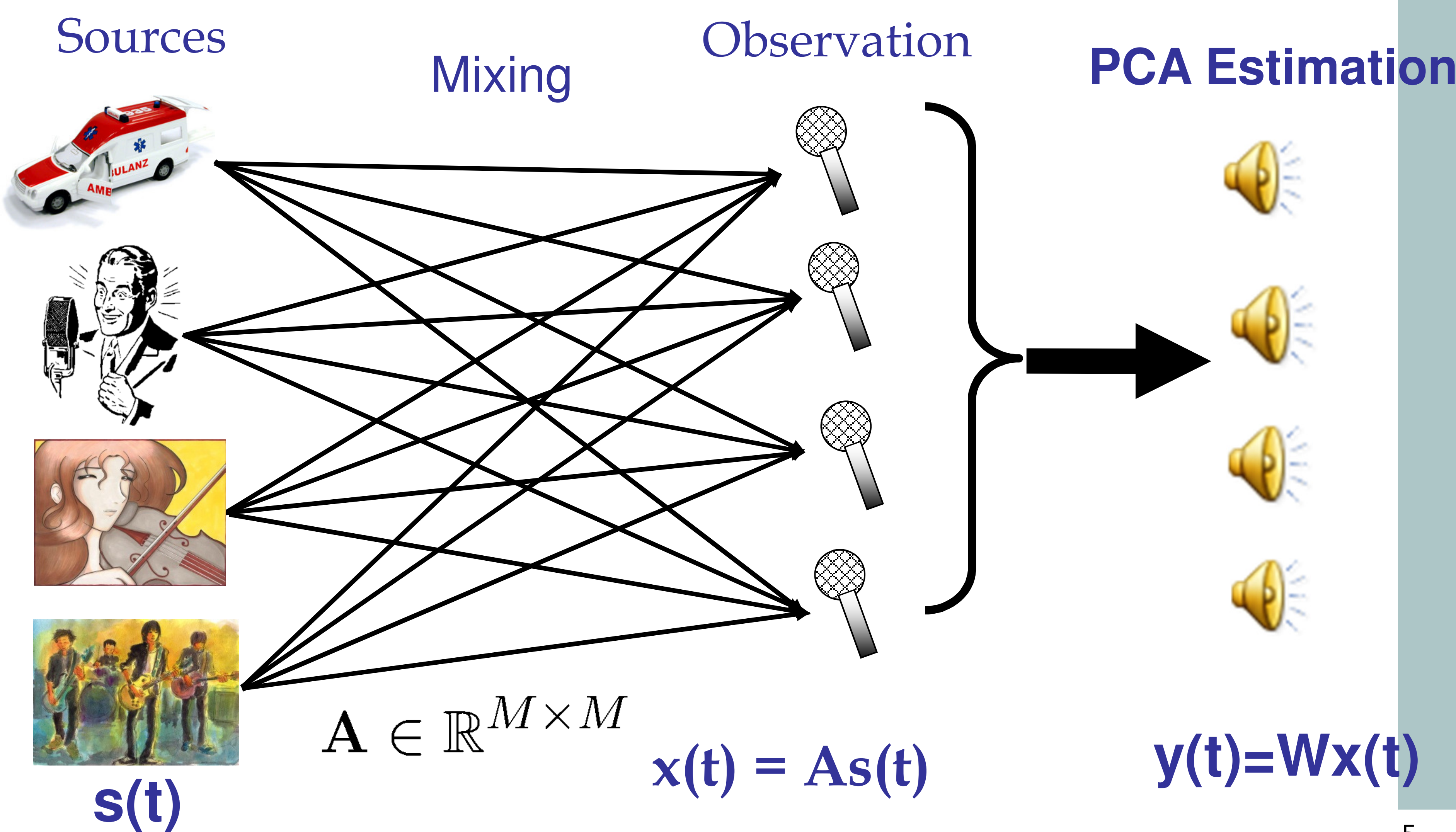
We want

$$\begin{pmatrix} s_1(1) \\ s_2(1) \end{pmatrix}, \begin{pmatrix} s_1(2) \\ s_2(2) \end{pmatrix}, \ldots, \begin{pmatrix} s_1(t) \\ s_2(t) \end{pmatrix}$$

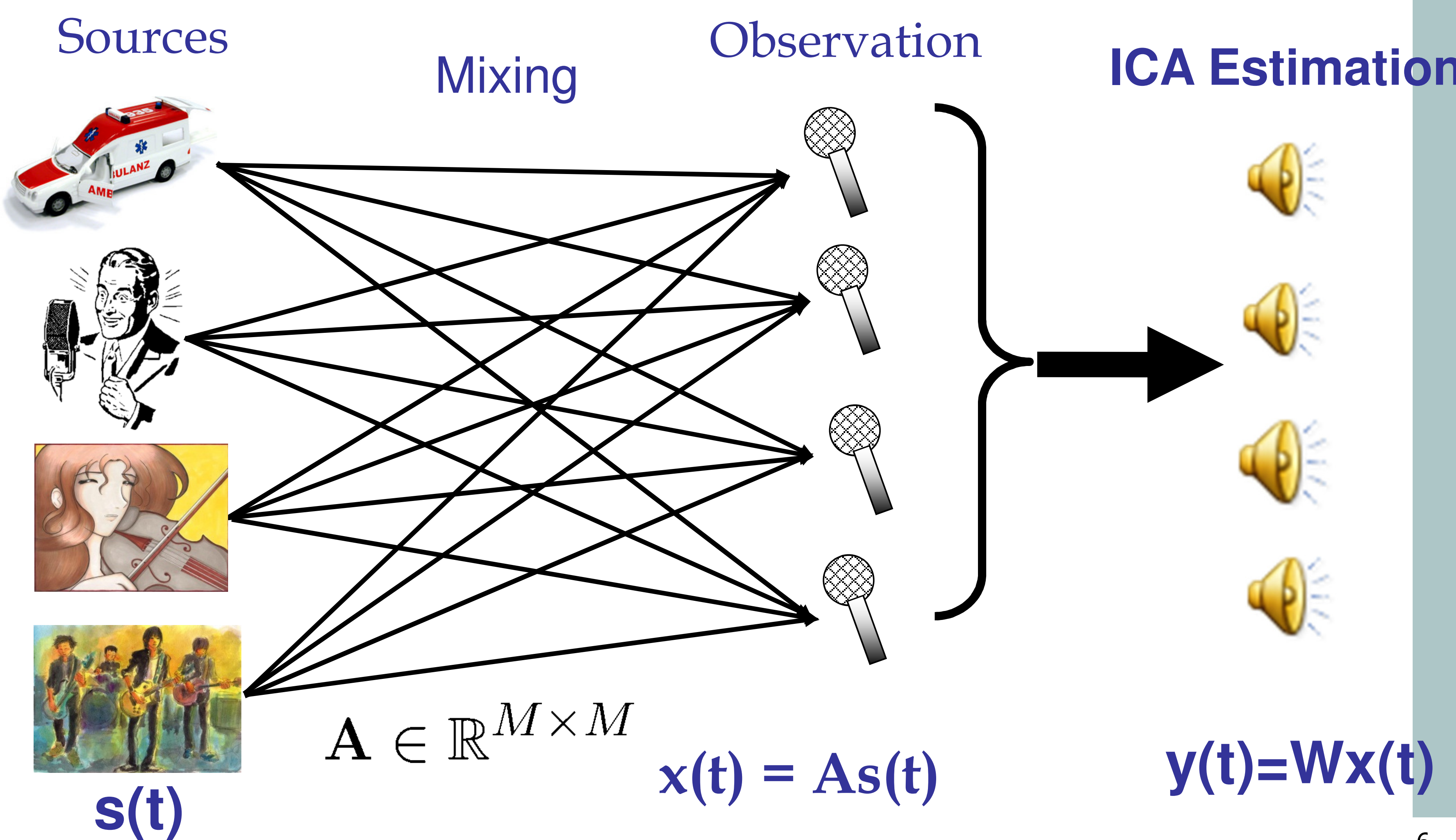
But we don't know $\{a_{ij}\}$, nor $\{s_i(t)\}$

Goal: Estimate $\{s_i(t)\}$, (and also $\{a_{ij}\}$)

The Cocktail Party Problem
SOLVING WITH PCA



The Cocktail Party Problem
SOLVING WITH ICA



ICA Theory

Statistical (in)dependence

Definition (Independence)

$Y_1,\ Y_2$ are independent $\Leftrightarrow p(y_1, y_2) \;=\; p(y_1)\,p(y_2)$

Definition (Shannon entropy)

$$H(\mathbf{Y}) \doteq H(Y_1, \ldots, Y_m) \doteq -\int p(y_1, \ldots, y_m) \log p(y_1, \ldots, y_m) d\mathbf{y}.$$

Definition (KL divergence)

$$0 \leq KL(f \| g) = \int f(x) \log \frac{f(x)}{g(x)}\, dx$$

Definition (Mutual Information)

$$0 \leq I(Y_1, \ldots, Y_M) \doteq \int p(y_1, \ldots, y_M) \log \frac{p(y_1, \ldots, y_M)}{p(y_1) \ldots p(y_M)} d\mathbf{y}$$

Solving the ICA problem with i.i.d. sources

ICA problem: $\mathbf{x} = \mathbf{A}\mathbf{s}$, $\mathbf{s} = [s_1; \ldots; s_M]$ are jointly independent.

Ambiguity:

$\mathbf{s} = [s_1; \ldots; s_M]$ sources can be recovered only up to

sign, scale and permutation.

Proof:

- $\mathbf{P}$ = arbitrary permutation matrix,
- $\mathbf{\Lambda}$ = arbitrary diagonal scaling matrix.

$$\Rightarrow \mathbf{x} = [\mathbf{A}\mathbf{P}^{-1}\mathbf{\Lambda}^{-1}][\mathbf{\Lambda}\mathbf{P}\mathbf{s}]$$

Solving the ICA problem

Lemma:

We can assume that $E[\mathbf{s}] = 0$.

Proof:

Removing the mean does not change the mixing matrix.

$$\mathbf{x} - E[\mathbf{x}] = \mathbf{A}(\mathbf{s} - E[\mathbf{s}]).$$

In what follows we assume that $E[\mathbf{s}\mathbf{s}^T] = \mathbf{I}_M$, $E[\mathbf{s}] = 0$.

Whitening

- Let $\Sigma \doteq cov(\mathbf{x}) = E[\mathbf{x}\mathbf{x}^T] = \mathbf{A}E[\mathbf{s}\mathbf{s}^T]\mathbf{A}^T = \mathbf{A}\mathbf{A}^T$.

  (We assumed centered data)

- Do **SVD**: $\Sigma \in \mathbb{R}^{N \times N}$, $rank(\Sigma) = M$,

  $\Rightarrow \Sigma = \mathbf{U}\mathbf{D}\mathbf{U}^T$,

  where $\mathbf{U} \in \mathbb{R}^{N \times M}$, $\mathbf{U}^T\mathbf{U} = \mathbf{I}_M$, **Signular vectors**

  $\mathbf{D} \in \mathbb{R}^{M \times M}$, diagonal with rank M. **Singular values**

Whitening

- Let $\mathbf{Q} \doteq \mathbf{D}^{-1/2}\mathbf{U}^T \in \mathbb{R}^{M \times N}$ *whitening matrix*
- Let $\mathbf{A}^* \doteq \mathbf{QA}$
- $\mathbf{x}^* \doteq \mathbf{Qx} = \mathbf{QAs} = \mathbf{A}^*\mathbf{s}$ is our new (*whitened*) ICA task.

We have,

$$E[x^* x^{*T}] = E[Qxx^T Q^T] = Q\Sigma Q^T = (D^{-1/2}U^T)UDU^T(UD^{-1/2}) = I_M$$

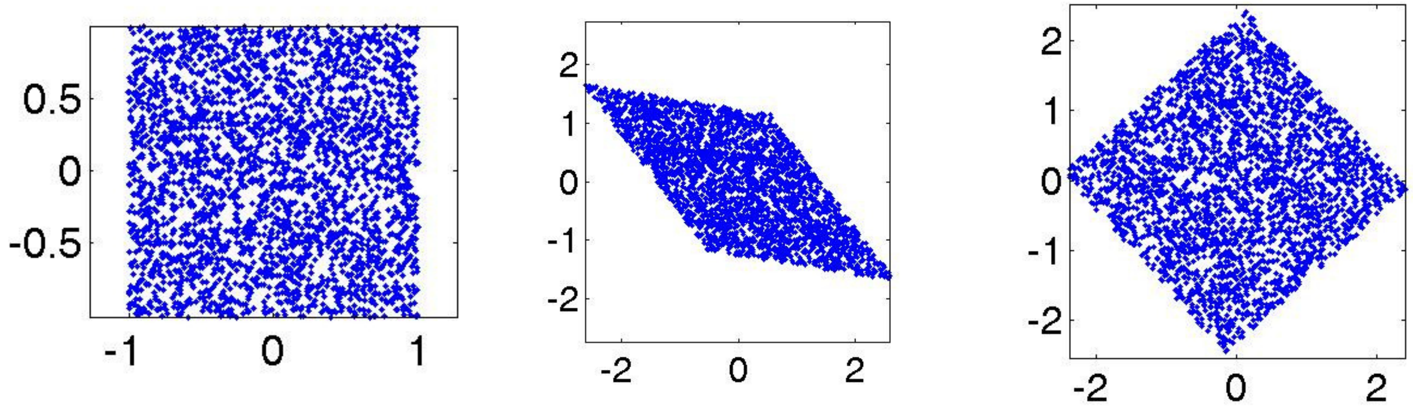
$$\boxed{\Rightarrow E[\mathbf{x}^*\mathbf{x}^{*T}] = \mathbf{I}_M, \text{ and } \mathbf{A}^*\mathbf{A}^{*T} = \mathbf{I}_M.}$$

Whitening solves half of the ICA problem

Note:

The number of free parameters of an N by N orthogonal matrix is (N-1)(N-2)/2.

$\Rightarrow$ whitening solves **half** of the ICA problem



original mixed whitened

After whitening it is enough to consider **orthogonal matrices** for separation.

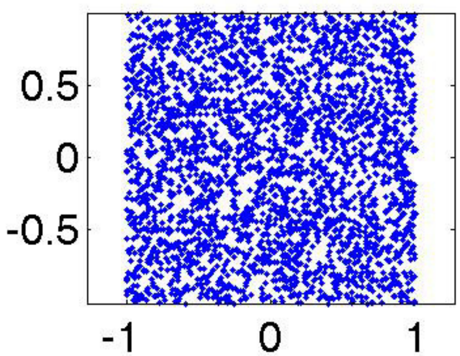
Solving ICA

ICA task: Given **x**,

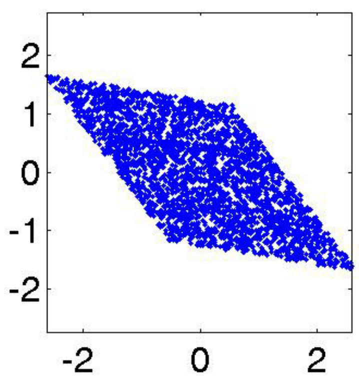
- ❑ find **y** (the estimation of **s**),
- ❑ find **W** (the estimation of $\mathbf{A}^{-1}$)

ICA solution: **y=Wx**

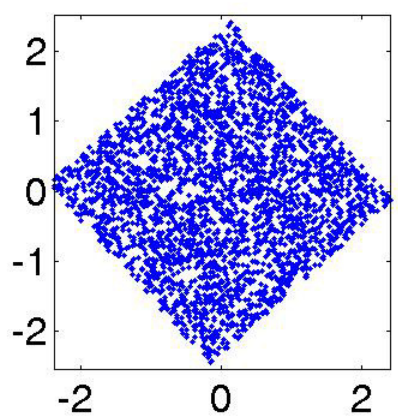
- ❑ Remove mean, E[**x**]=0
- ❑ Whitening, E[$\mathbf{x}\mathbf{x}^T$]=**I**
- ❑ Find an orthogonal **W** optimizing an objective function
  - Sequence of 2-d Jacobi (Givens) rotations



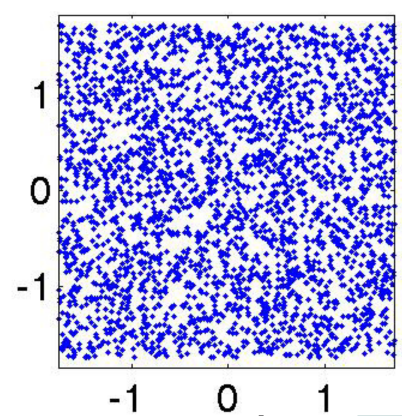
original



mixed



whitened



rotated

(demixed)

Optimization Using Jacobi Rotation Matrices

$$\mathbf{G}(p,q,\theta) \doteq \begin{array}{cc} \begin{pmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & \cos(\theta) & \cdots & -\sin(\theta) & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & \sin(\theta) & \cdots & \cos(\theta) & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{pmatrix} & \begin{array}{l} \\ \\ \leftarrow \mathrm{p} \\ \\ \leftarrow \mathrm{q} \\ \\ \\ \end{array} \\ \begin{array}{ccccccc} & & \uparrow & & \uparrow & & \\ & & \mathrm{p} & & \mathrm{q} & & \end{array} & \end{array} \in \mathbf{R}^{M\times M}$$

$$\textit{Observation}: \mathbf{x} = \mathbf{A}\mathbf{s}$$

$$\textit{Estimation}: \mathbf{y} = \mathbf{W}\mathbf{x}$$

$$\mathbf{W} = \arg \min_{\tilde{\mathbf{W}} \in \mathcal{W}} J(\tilde{\mathbf{W}}\mathbf{x}),$$

$$\text{where } \mathcal{W} = \{\mathbf{W} | \mathbf{W} = \prod_i G(p_i, q_i, \theta_i)\}$$

ICA Cost Functions

Let $\mathbf{y} \doteq \mathbf{W}\mathbf{x}$, $\mathbf{y} = [y_1; \ldots; y_M]$, and let us measure the dependence using Shannon's mututal information:

$$J_{ICA_1}(\mathbf{W}) \doteq I(y_1, \ldots, y_M) \doteq \int p(y_1, \ldots, y_M) \log \frac{p(y_1,\ldots,y_M)}{p(y_1)\ldots p(y_M)} d\mathbf{y},$$

Let $H(\mathbf{y}) \doteq H(y_1, \ldots, y_m) \doteq -\int p(y_1, \ldots, y_m) \log p(y_1, \ldots, y_m) d\mathbf{y}$.

Lemma

$$H(\mathbf{W}\mathbf{x}) = H(\mathbf{x}) + \log |\det \mathbf{W}|$$

Proof: Homework

Therefore,

$$\begin{aligned} I(y_1, \ldots, y_M) &= \int p(y_1, \ldots, y_M) \log \frac{p(y_1, \ldots, y_M)}{p(y_1) \ldots p(y_M)} \\ &= -H(y_1, \ldots, y_M) + H(y_1) + \ldots + H(y_M) \\ &= -H(x_1, \ldots, x_M) - \log |\det \mathbf{W}| + H(y_1) + \ldots + H(y_M). \end{aligned}$$

ICA Cost Functions

$$
\begin{aligned}
I(y_1,\ldots,y_M) &= \int p(y_1,\ldots,y_M)\log\frac{p(y_1,\ldots,y_M)}{p(y_1)\ldots p(y_M)} \\
&= -H(y_1,\ldots,y_M) + H(y_1) + \ldots + H(y_M) \\
&= -H(x_1,\ldots,x_M) - \log|\det \mathbf{W}| + H(y_1) + \ldots + H(y_M).
\end{aligned}
$$

$H(x_1,\ldots,x_M)$ is constant, $\log|\det \mathbf{W}| = 0$.

Therefore,

$$\boxed{J_{ICA_2}(\mathbf{W}) \doteq H(y_1) + \ldots + H(y_M)}$$

The covariance is fixed: I. Which distribution has the largest entropy?

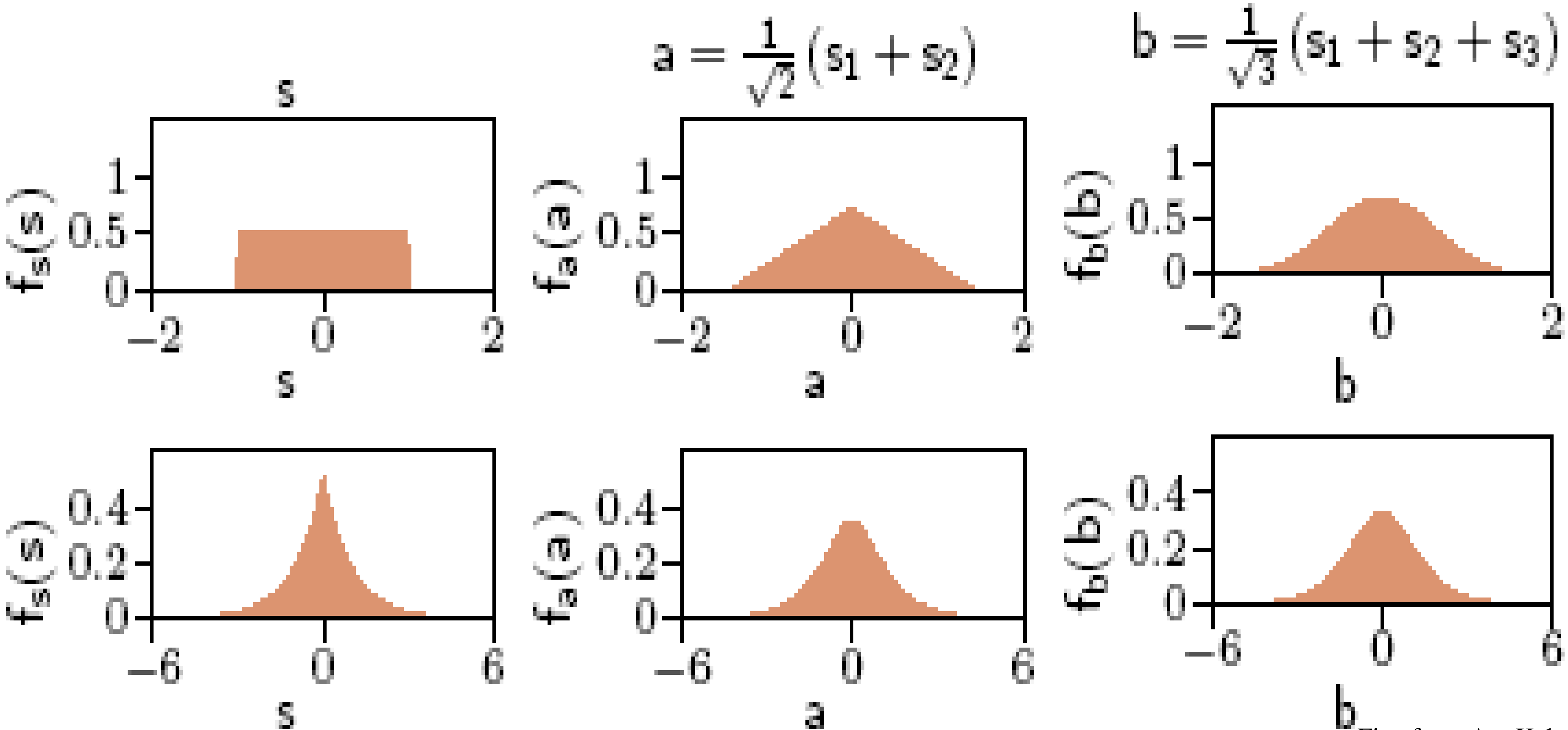
$\Rightarrow$ go away from normal distribution

Central Limit Theorem

The sum of independent variables converges to the normal distribution

$\Rightarrow$ For separation go far away from the normal distribution

$\Rightarrow$ **Negentropy, |kurtozis| maximization**



Figs from Ata Kaban

ICA Algorithms

Maximum Likelihood ICA Algorithm

David J.C. MacKay (97)

- simplest approach
- requires knowing densities of hidden sources $\{f_i\}$

rows of **W**

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t),\ \mathbf{s}(t) = \mathbf{W}\mathbf{x}(t),\ \text{where } \mathbf{A}^{-1} = \mathbf{W} = [\mathbf{w}_1; \ldots; \mathbf{w}_M] \in \mathbb{R}^{M\times M}$$

$$L = \sum_{t=1}^{T} \log p_x(x(t)) = \sum_{t=1}^{T} \log \underbrace{p_x(As(t))}_{A^{-1} p_s(\underbrace{A^{-1} x(t)}_{s(t)})} \Rightarrow \max_{A}$$

$$p_{As}(u) = A^{-1} p_s(A^{-1} u)$$

$$\Rightarrow L = \sum_{t=1}^{T} \log A^{-1} p_s(s(t)) = T \log|W| + \sum_{t=1}^{T} \log \underbrace{p_s(s(t))}_{\prod_{i=1}^{M} p_{s_i}(s_i(t))}$$

$$= T \log|W| + \sum_{t=1}^{T} \sum_{i=1}^{M} \log \underbrace{p_{s_i}(w_i x(t))}_{f_i(w_i x(t))}$$

$$\Rightarrow \max_{W}$$

Maximum Likelihood ICA Algorithm

$$L = T \log|W| + \sum_{t=1}^{T} \sum_{k=1}^{M} \log f_k(w_k x(t))$$

$$\Rightarrow \max_{W} L \Rightarrow \frac{\partial L}{\partial w_{ij}} = ?$$

$$\frac{\partial L}{\partial w_{ij}} = T(W^T)^{-1}_{ij} + \sum_{t=1}^{T} \underbrace{\frac{\partial}{\partial w_{ij}} \sum_{k=1}^{M} \log f_k(w_k x(t))}_{\frac{f_i'(w_i x(t))}{f_i(w_i x(t))} x_j(t)}$$

$$= T(W^T)^{-1}_{ij} + \sum_{t=1}^{T} \frac{f_i'(w_i x(t))}{f_i(w_i x(t))} x_j(t)$$

$$\Rightarrow \boxed{\Delta \mathbf{W} \propto [\mathbf{W}^T]^{-1} + \frac{1}{T} \sum_{t=1}^{T} g(\mathbf{W}\mathbf{x}(t))\mathbf{x}^T(t),}$$ where $g_i = f_i'/f_i$

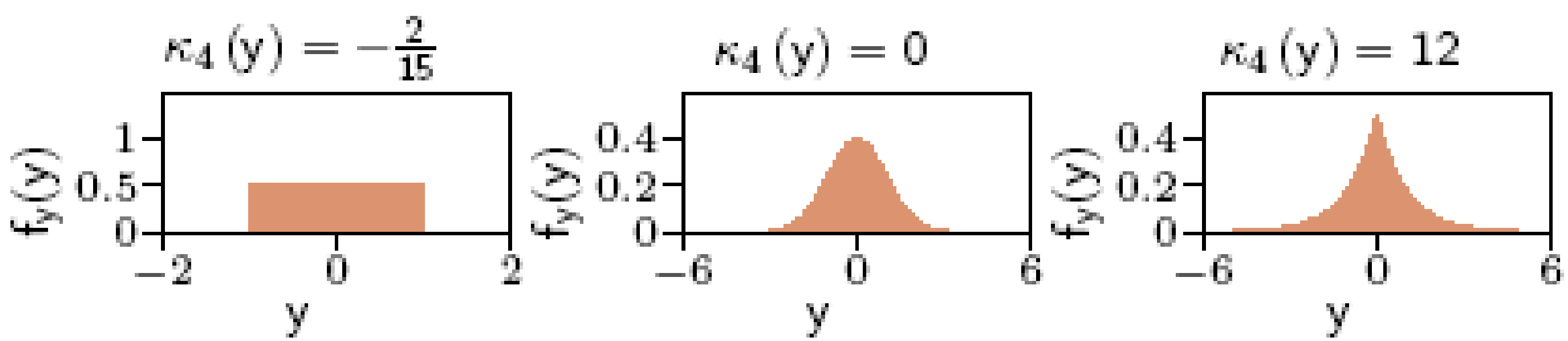
ICA algorithm based on Kurtosis maximization

Kurtosis = 4th order cumulant

Measures

- the distance from normality
- the degree of peakedness

- $$\kappa_4(y) = \mathrm{E}\{y^4\} - \underbrace{3\left(\mathrm{E}\{y^2\}\right)^2}_{= 3 \text{ if } \mathrm{E}\{y\} = 0 \text{ and whitened}}$$



The Fast ICA algorithm (Hyvarinen)

- Given whitened data $\mathbf{z}$
- Estimate the 1^{st} ICA component:

Probably the most famous ICA algorithm

$\star\ y = \mathbf{w}^T\mathbf{z},\ \|\mathbf{w}\| = 1,$ $\quad \Leftarrow \mathbf{w}^T = 1^{st}$ row of $\mathbf{W}$

$$\boxed{\begin{array}{l}\star \text{ maximize kurtosis } f(\mathbf{w}) \doteq \kappa_4(y) \doteq \mathbb{E}[y^4]\text{-}3 \\ \qquad\qquad \text{with constraint } h(\mathbf{w}) = \|\mathbf{w}\|^2 - 1 = 0\end{array}}$$

$\star$ At optimum $f'(\mathbf{w}) + \lambda h'(\mathbf{w}) = \mathbf{0}^T$ $\quad$ (λ Lagrange multiplier)

$$\Rightarrow 4\mathbb{E}[(\mathbf{w}^T\mathbf{z})^3\mathbf{z}] + 2\lambda\mathbf{w} = 0$$

Solve this equation by Newton–Raphson's method.

The Fast ICA algorithm (Hyvarinen)

Solve: $F(\mathrm{w}) = 4\mathbb{E}[(\mathbf{w}^T\mathbf{z})^3\mathbf{z}] + 2\lambda\mathbf{w} = 0$

Note:

$$y = \mathbf{w}^T\mathbf{z},\ \|\mathbf{w}\| = 1,\ \mathbf{z} \text{ white} \Rightarrow \mathbb{E}[(\mathbf{w}^T\mathbf{z})^2] = 1$$

$\mathbb{E}[z z^T] = I$

$\mathbb{E}[w^T z z^T w] = w^T I w = 1$

$\|w\| = 1$

The derivative of F:

$$F'(\mathrm{w}) = 12\mathbb{E}[(\mathbf{w}^T\mathbf{z})^2\mathbf{z}\mathbf{z}^T] + 2\lambda\mathbf{I}$$
$$\sim 12\mathbb{E}[(\mathbf{w}^T\mathbf{z})^2]\mathbb{E}[\mathbf{z}\mathbf{z}^T] + 2\lambda\mathbf{I}$$
$$= 12\mathbb{E}[(\mathbf{w}^T\mathbf{z})^2]\mathbf{I} + 2\lambda\mathbf{I}$$
$$= 12\mathbf{I} + 2\lambda\mathbf{I}$$

The Fast ICA algorithm (Hyvarinen)

The Jacobian matrix becomes diagonal, and can easily be inverted.

$$\mathbf{w}(k+1) = \mathbf{w}(k) - [F'(\mathbf{w}(k)]^{-1} F(\mathbf{w}(k))$$

$$\mathbf{w}(k+1) = \mathbf{w}(k) - \frac{4\mathbb{E}[(\mathbf{w}(k)^T\mathbf{z})^3\mathbf{z}]+2\lambda\mathbf{w}(k)}{12+2\lambda}$$

$$(12+2\lambda)\mathbf{w}(k+1) = (12+2\lambda)\mathbf{w}(k) - 4\mathbb{E}[(\mathbf{w}(k)^T\mathbf{z})^3\mathbf{z}] - 2\lambda\mathbf{w}(k)$$

$$-\frac{12+2\lambda}{4}\mathbf{w}(k+1) = -3\mathbf{w}(k) + \mathbb{E}[(\mathbf{w}(k)^T\mathbf{z})^3\mathbf{z}]$$

Therefore,

Let $\mathbf{w}_1$ be the fix pont of:

$$\tilde{\mathbf{w}}(k+1) = \mathbb{E}[(\mathbf{w}(k)^T\mathbf{z})^3\mathbf{z}] - 3\mathbf{w}(k)$$
$$\mathbf{w}(k+1) = \frac{\tilde{\mathbf{w}}(k+1)}{\|\tilde{\mathbf{w}}(k+1)\|}$$

- Estimate the 2^{nd} ICA component similarly using the $\mathbf{w} \perp \mathbf{w}_1$ additional constraint... and so on ...

Other Nonlinearities

$$\max_{w} \mathbb{E}\, G(w^T z) \quad \text{s.t. } w^T w = 1$$

$G(y) = y^4 - 3$ ← IN THE PREVIOUS

$$\Rightarrow F(w) = \mathbb{E}\, z\, g(w^T z) - \lambda w = 0 \in \mathbb{R}^n \qquad g(y) = G'(y) \downarrow = 4y^3 \Downarrow g'(y) = 12y^2$$

$$\nabla F(w) = \mathbb{E}\left[z z^T g'(w^T z)\right] - \lambda I$$

$$\approx \mathbb{E}[z z^T]\, \mathbb{E}[g'(w^T z)] - \lambda I$$

$$= \underbrace{\mathbb{E}[g'(w^T z)]}_{\text{SCALAR}} I - \lambda I$$

↑

DIAGONAL MTX WITH IDENTICAL ELEMENTS

Other Nonlinearities

Newton method:

$$w(k+1) = w(k) - \left[\nabla F(w(k))\right]^{-1} F(w(k))$$

$$= w(k) - \frac{\mathbb{E}\left[z\, g(w_{(k)}^T z)\right] - \alpha w(k)}{\mathbb{E}\left[g'(w_{(k)}^T z)\right] - \alpha}$$

$$\Rightarrow \left(\mathbb{E}\left[g'(w_{(k)}^T z)\right] - \alpha\right) w(k+1) = \left(\mathbb{E}\left[g'(w_{(k)}^T z)\right] - \alpha\right) w(k) + \alpha w(k) - \mathbb{E}\left[z\, g(w(k)^T z)\right]$$

$$\underset{c \in \mathbb{R}}{\uparrow c}\; w(k+1) = \mathbb{E}\left[z\, g(w(k)^T z)\right] - \mathbb{E}\left[g'(w(k)^T z)\right] w(k)$$

Algorithm:

$$\Rightarrow w(k+1) = \mathbb{E}\left[z\, g(w(k)^T z)\right] - \mathbb{E}\left[g'(w(k)^T z)\right] w(k)$$

$$w(k+1) = w(k+1) / \| w(k+1) \|$$

$$g_1(u) = \mathrm{TANH}(a_1 u)$$

$$g_2(u) = u\, \mathrm{EXP}(-u^2/2)$$

Fast ICA for several units

$$W = \begin{bmatrix} w_1^T \\ \vdots \\ w_n^T \end{bmatrix} \qquad WW^T = I$$

WE NEED TO PREVENT DIFFERENT VECTORS CONVERGING TO THE SAME MAXIMA

IF WE ALREADY ESTIMATED $w_1 \ldots w_p$

THEN UPDATE w_{p+1}

AND AFTER THAT

$$w_{p+1} = w_{p+1} - \sum_{j=1}^{p} (w_{p+1}^T w_j)\, w_j$$

GRAM-SCHMIDT DECORRELATION

OPTION 2:

$$W = (WW^T)^{-1/2}\, W$$

SYMMETRIC DECORRELATION

FROM SVD $WW^T = UDU^T$

$$(WW^T)^{-1/2} = UD^{-1/2}U^T$$